1 Inductive Bias

less for transformer

2 Weight Reuse and Sharing

1. FC
2. Conv: Sharing in space. Differs when rotating. so rand-rotating req
3. Recurrent: Sharing in time. “Hidden state”

3 transformer

3.1 tokenize

token set? not exhaustive
character set, not enough
subword :bert: word piece, by pair encoding

3.2 encoder

Position encoding embedding in context: transformer encoding.

3.3 decoder

masked: cannot foresee. Query: from encoder to decoder. project

3.4 add&normalize

prenorm (we get it, stable) / postnorm (warmup)

4 VIT:Vision transformer

Patch, linear projection (conv, lol)
class token -> transformer encoder -> MLP head (prenorm)

5 attention

B C H W -> B C N -> B N C -> B N 3C (qkv) -> 3 B N C D
B H N D -> B H N N -> scale $\frac{1}{\sqrt{h}}$